Questions

Q1.

A lake contains three different types of carp.

There are an estimated 450 mirror carp, 300 leather carp and 850 common carp.

Tim wishes to investigate the health of the fish in the lake.

He decides to take a sample of 160 fish.

(a) Give a reason why stratified random sampling cannot be used.

(1)

(b) Explain how a sample of size 160 could be taken to ensure that the estimated populations of each

type of carp are fairly represented.
You should state the name of the sampling method used.

(2)

As part of the health check, Tim weighed the fish.

His results are given in the table below.

Weight (w kg)	Frequency (f)	Midpoint (m kg)
$2 \leqslant w < 3.5$	8	2.75
$3.5 \leqslant w < 4$	32	3.75
$4 \leqslant w < 4.5$	64	4.25
$4.5 \leqslant w < 5$	40	4.75
$5 \leqslant w < 6$	16	5.5

(You may use $\sum fm = 692$ and $\sum fm^2 = 3053$)

(c) Calculate an estimate for the standard deviation of the weight of the carp.

(2)

Tim realised that he had transposed the figures for 2 of the weights of the fish.

He had recorded in the table 2.3 instead of 3.2 and 4.6 instead of 6.4.

(d) Without calculating a new estimate for the standard deviation, state what effect

(i) using the correct figure of 3.2 instead of 2.3
(ii) using the correct figure of 6.4 instead of 4.6
would have on your estimated standard deviation.
Give a reason for each of your answers.

(2)

(Total for question = 7 marks)

Q2.

Helen is studying one of the qualitative variables from the large data set for Heathrow from 2015.

She started with the data from 3rd May and then took every 10th reading.

There were only 3 different outcomes with the following frequencies

Outcome	*A*	*B*	*C*
Frequency	16	2	1

(a) State the sampling technique Helen used.

(1)

(b) From your knowledge of the large data set

(i) suggest which variable was being studied,
(ii) state the name of outcome *A*.

(2)

George is also studying the same variable from the large data set for Heathrow from 2015. He started with the data from 5th May and then took every 10th reading and obtained the following

Outcome	*A*	*B*	*C*
Frequency	16	1	1

Helen and George decided they should examine all of the data for this variable for Heathrow from 2015 and obtained the following

Outcome	*A*	*B*	*C*
Frequency	155	26	3

(c) State what inference Helen and George could reliably make from their original samples about the outcomes of this variable at Heathrow, for the period covered by the large data set in 2015.

(1)

(Total for question = 4 marks)

Q3.

Charlie is studying the time it takes members of his company to travel to the office.
He stands by the door to the office from 08 40 to 08 50 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

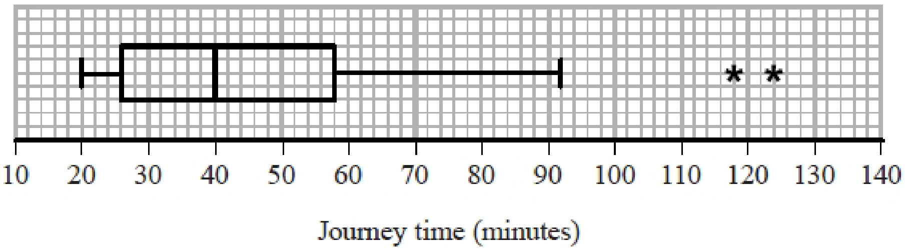
(2)

Taruni decided to ask every member of the company the time, *x* minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Taruni's results are summarised by the box plot and summary statistics below.



$$n = 95 \quad \sum x = 4133 \quad \sum x^2 = 202\,294$$

(d) Write down the interquartile range for these data.

(1)

(e) Calculate the mean and the standard deviation for these data.

(3)

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

Rana and David both work for the company and have both moved house since Taruni collected her data.

Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

(Total for question = 13 marks)

Q4.

(a) State one disadvantage of using quota sampling compared with simple random sampling.

(1)

In a university 8% of students are members of the university dance club.

A random sample of 36 students is taken from the university.

The random variable X represents the number of these students who are members of the dance club.

(b) Using a suitable model for X, find

(i) $\mathrm{P}(X = 4)$
(ii) $\mathrm{P}(X \geq 7)$

(3)

Only 40% of the university dance club members can dance the tango.

(c) Find the probability that a student is a member of the university dance club and can dance the tango.

(1)

A random sample of 50 students is taken from the university.

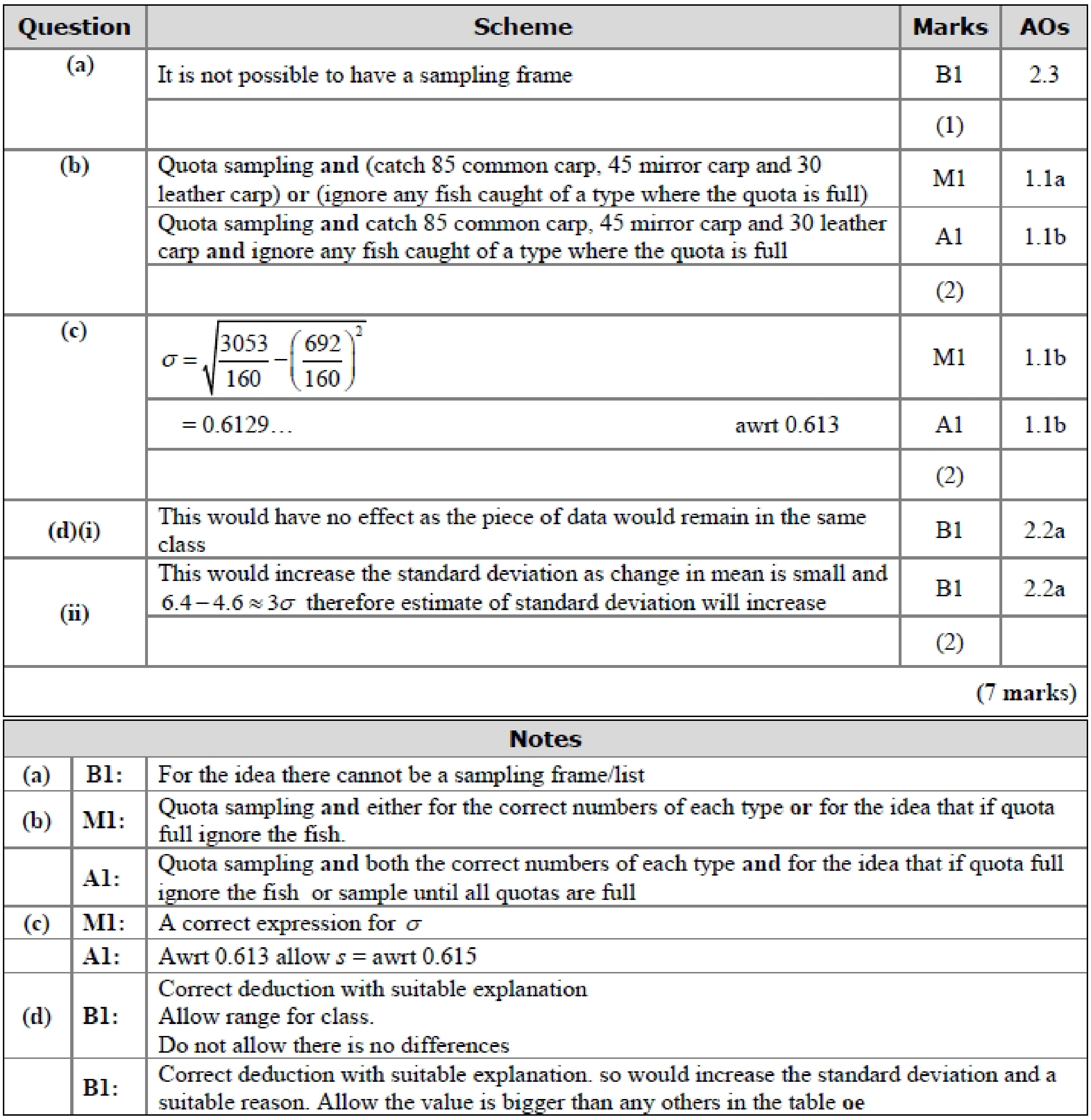
(d) Find the probability that fewer than 3 of these students are members of the university dance club and can dance the tango.

(2)

(Total for question = 7 marks)

Mark Scheme

Q1.



Question	Scheme	Marks	AOs
(a)	It is not possible to have a sampling frame	B1	2.3
		(1)	
(b)	Quota sampling **and** (catch 85 common carp, 45 mirror carp and 30 leather carp) **or** (ignore any fish caught of a type where the quota is full)	M1	1.1a
	Quota sampling **and** catch 85 common carp, 45 mirror carp and 30 leather carp **and** ignore any fish caught of a type where the quota is full	A1	1.1b
		(2)	
(c)	$\sigma = \sqrt{\frac{3053}{160} - \left(\frac{692}{160}\right)^2}$	M1	1.1b
	$= 0.6129\ldots$ awrt 0.613	A1	1.1b
		(2)	
(d)(i)	This would have no effect as the piece of data would remain in the same class	B1	2.2a
(ii)	This would increase the standard deviation as change in mean is small and $6.4 - 4.6 \approx 3\sigma$ therefore estimate of standard deviation will increase	B1	2.2a
		(2)	
			(7 marks)

Notes		
(a)	B1:	For the idea there cannot be a sampling frame/list
(b)	M1:	Quota sampling **and** either for the correct numbers of each type **or** for the idea that if quota full ignore the fish.
	A1:	Quota sampling **and** both the correct numbers of each type **and** for the idea that if quota full ignore the fish or sample until all quotas are full
(c)	M1:	A correct expression for σ
	A1:	Awrt 0.613 allow $s =$ awrt 0.615
(d)	B1:	Correct deduction with suitable explanation Allow range for class. Do not allow there is no differences
	B1:	Correct deduction with suitable explanation. so would increase the standard deviation and a suitable reason. Allow the value is bigger than any others in the table **oe**

Q2.

Qu	Scheme	Marks	AO
(a)	Systematic (sampling)	B1	1.2
		(1)	
(b)(i)	[Daily Mean] Wind Speed	B1	2.2a
(ii)	Light	B1	1.2
		(2)	
(c)	Variable *A* occurs most (around 80–90%) of the time	B1	2.2b
		(1)	
		(4 marks)	

	Notes
(a)	B1 for identifying the correct sampling technique Allow slight misspelling e.g. "sysmatic", "sytmatic" Do NOT allow "systemic"
(b)(i)	B1 for identifying appropriate qualitative variable. {LDS mark} Allow "Wind speed" or "Wind strength" but NOT just "wind" or "wind direction"
(ii)	B1 for realising that modal wind speed is "Light" {LDS mark} Allow just "light" or "most light"
NB	These two B marks are independent so can score B0B1 for e.g. "rainfall" and "light"
(c)	B1 for inferring that frequency of *A* can be estimated fairly reliably: {underestimates *B* and over estimates *C*} e.g. "*A* is the most frequent" [can then ignore comments about *B* and *C*]

Q3.

Qu	Scheme	Marks	AO
(a)	Convenience or opportunity [sampling]	B1	1.2
		(1)	
(b)	Quota [sampling]	B1	1.1a
	e.g. Take 4 people every 10 minutes	B1	1.1b
		(2)	
(c)	Census	B1	1.2
		(1)	
(d)	[58 – 26 =] **32** (min)	B1	1.1b
		(1)	
(e)	$\mu = \frac{4133}{95} = 43.505263\ldots$ awrt **43.5 (min)**	B1	1.1b
	$\sigma_x = \sqrt{\frac{202\,294}{95} - \mu^2} = \sqrt{236.7026\ldots}$	M1	1.1b
	$= 15.385\ldots$ **awrt 15.4 (min)**	A1	1.1b
		(3)	
(f)	There are outliers in the data (or data is skew) which will affect mean and sd	B1	2.4
	Therefore use median and IQR	dB1	2.4
		(2)	
(g)	Value of 20, LQ at 26 and outliers will not change or state that median and upper quartile are the values that do change	B1	1.1b
	More values now below 40 than above so Q_2 or Q_3 will change and be lower	M1	2.1
	Both Q_2 and Q_3 will be lower	A1	2.4
		(3)	
		(13 marks)	

	Notes
(b)	1st B1 for quota (sampling) mentioned ("Stratified" or "systematic" or "random" are B0B0) 2nd B1 for a description of how such a system might work, requires suitable strata or categories e.g. time slots, departments, gender, age groups, distance travelled etc Suggestion of randomness is B0
(e)	B1 for a correct mean (awrt 43.5) M1 for a correct expression for the sd (including $\sqrt{\ }$)ft their mean A1 for awrt 15.4 (Allow s = 15.4667… awrt 15.5)
(f)	1st B1 for acknowledging outliers or skewness are a problem for mean and sd "extreme values"/"anomalies" OK May be implied by saying median and IQR not affected by.. We need to see mention of "outliers", "skewness" and the problem so "data is skewed so use median and IQR" is B0 unless mention that they are not affected by extreme values or mean and standard deviation can be "inflated" by the positive skew etc 2nd dB1 dep on 1st B1 for therefore choosing median and IQR
(g)	B1 for identifying 2 of these 3 groups of unchanged values or stating only Q_2 and Q_3 change M1 for explaining that median or UQ should be lower. E.g. the 2 values have moved to below 40 (or 58) and therefore more than 50% below 40 or (more than 75% below 58) or an argument to show that the other 3 values are the same. (o.e.) Allow arrows on box plot provided statement in words about increased % below 40 or 58 etc A1 for stating median and UQ are both lower with clear evidence of M1 scored [If lots of values on 40 then median might not change but, since two values do change then UQ would change. If this meant that 92 became an outlier then we would have a new value for upper whisker and an extra outlier so effectively 3 values are altered. So median changes]

Q4.

	Scheme	Marks	AO
(a)	**Disadvantage:** e.g. Not random; cannot use (reliably) for inferences	B1	1.1b
		(1)	
(b)	[Sight or correct use of] $X \sim B(36, 0.08)$	M1	3.3
(i)	$P(X = 4) = 0.167387\ldots$ awrt $\underline{0.167}$	A1	1.1b
(ii)	[P(*X* …7) = 1 − P(*X* „ 6) =] $0.022233\ldots$ awrt $\underline{0.0222}$	A1	1.1b
		(3)	
(c)	P(In dance club and dance tango) = $0.4 \times 0.08 = \underline{0.032}$ or $\underline{\underline{\frac{4}{125}}}$ or $\underline{3.2\%}$	B1	1.1b
		(1)	
(d)	[Let *T* = those who can dance the Tango. Sight or use of] $T \sim B(50, \text{"}0.032\text{"})$	M1	3.3
	[P(*T* < 3) = P(*T* „ 2) =] $0.7850815\ldots$ awrt $\underline{0.785}$	A1	1.1b
		(2)	
		(7 marks)	

	Notes
(a)	B1 for a suitable disadvantage:

Allow (B1)	Do NOT allow (B0)
Not random $\underline{\text{or}}$ less random (o.e.)	Not representative
Cannot use (reliably) for inferences	Less accurate
(More likely to be) biased	Any comment based on time or cost
	Any mention of skew
	Any mention of non-response

(b)	M1 for sight of B(36, 0.08) Allow in words: $\underline{\text{binomial}}$ with $\underline{n = 36}$ and $\underline{p = 0.08}$ may be implied by one correct answer to 2sf $\underline{\text{or}}$ sight of P(*X* „ 6) = 0.97776... i.e. awrt 0.98 Allow for $36C4 \times 0.08^4 \times 0.92^{32}$ as this is "correct use"
(i)	1st A1 for awrt 0.167 NB An answer of just awrt 0.167 scores M1(⇒)1st A1
(ii)	2nd A1 for awrt 0.0222
(c)	B1 for 0.032 o.e. (Can allow for sight of 0.4×0.08)
(d)	M1 for sight of B(50, "0.032") ft their answer to (c) provided it is a probability ≠ 0.08 may be implied by correct answer $\underline{\text{or}}$ sight of [P(*T* „ 3)] = 0.924348…i.e. awrt 0.924 or P(*T* „ 2) as part of 1 − P(*T* „ 2) calc. A1 for awrt 0.785
MR	Allow MR of 50 (e.g. 30) provided clearly attempting P(*T* „ 2) and score M1A0